COURSE WORK 1: DATA MINING

MSc Data Science: Coventry University UK (**2024.2 BATCH**)

Name: P D Lolitha Lakshan Weerasinghe

Index Number: COMScDS242P-003

Question 1: Online Retail Data

GitHub Link - Code: <https://github.com/lolitha-lakshan-1/DataMining_assignment_1/blob/main/Assignment.Rmd>

## Task 1 – Data Understanding

As the initial task excel dataset was loaded to R Studio for preliminary analysis. In this step dataset was checked for,

1. How the dataset composed, what are the columns which exists?
   1. It was found all the said columns existed as mentioned which are,

InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country

1. Are the column data types being correct?
   1. It was noted that most column types contained correct data types including the “InvoiceDate”. However, it was noted “Stock Code” contained alpha numeric values rather than a 5-digit integral number.
2. What are the missing values?
   1. It was found that “Description” value is missing from 1454 transactions and “CustomerID” is missing from 135080 transactions.
3. What are the unique values in columns such as “Country” and “Stock Code”
   1. “Country” column contained all the valid country values except there were some instances of value being “Unspecified”. There are 446 transactions with country specified as “Unspecified.”
   2. “Stock Code” contained non product codes such as 'BANK CHARGES', 'POST', 'DOT', 'M', 'PADS', 'C2'
4. Are there any outliers in the numeric columns?
   1. “Quantity” column contained values ranging from -80995.00 to 80995.00. However, the mean is 9.55 and median being 3.00. There are 10624 transactions with negative quantities.
   2. “Unit Price” ranges from -11062.06 to 38970.00 while mean being 4.61 and median being 2.08. There are 2 transactions with negative unit prices.

Box Plot of Quantity prior to removing outliers.

A black and white diagram of a number

AI-generated content may be incorrect.

Box Plot of Unit Price prior to removing outliers.

A graph of a number of numbers

AI-generated content may be incorrect.

1. Are there any correlated columns?

A green squares with white text

AI-generated content may be incorrect.

According to the heatmap there were not feature correlations found.

### Data Cleansing

1. Removed transactions which contained negative value for Unit Price or Quantity fields. About 10626 transactions were removed.
2. Removed cancelled transactions which contained “C” in the InvoiceNo.
3. Removed transactions with non-product Stock Codes such as 'BANK CHARGES', 'POST', 'DOT', 'M', 'PADS', 'C2'. About 2324 transactions were removed from the dataset.
4. Removed transactions with Country specified as “Unspecified”. About 446 transactions were removed.
5. It was decided not to remove customer null transactions as it does not affect the basket values.
6. After cleansing there are 528513 transactions remaining on the dataset.

## Task 2 – Apriori Implementation

After the data cleansing dataset was grouped using the combination of Country and the InvoiceNo. This would result unique combinations of Country and InvoiceNo, removing any duplicates. Then it will count the number of distinct InvoiceNo per country, and sort them in descending order of count and would output the top three countries with the number of transactions.

top\_countries <- data %>%

distinct(Country, InvoiceNo) %>%

count(Country, sort = TRUE) %>%

slice\_head(n = 3) %>%

pull(Country)

Countries with most transaction counts:

1. United Kingdom 23494
2. Germany 603
3. France 461

Thereafter, transaction list was creating using the StockCode and InvoiceNo from the top three countries.

Following code filters the dataset for a top country and keeps only unique combinations of InvoiceNo and StockCode to avoid duplicate items in the same transaction.

It then groups the items by invoice number to form a Apriori transaction list, which is converted into a transactions object required for Apriori analysis.

The Apriori algorithm is run with minimum support of 1%, confidence of 50%, and at least 2 items per rule. The resulting rules are sorted by lift to highlight the strongest associations.

for (country in top\_countries) {

cat("\n=====================\n")

cat("Running Apriori for:", country, "\n")

cat("=====================\n")

# Filter for country

country\_data <- data %>%

filter(Country == country) %>%

distinct(InvoiceNo, StockCode)

# Convert to transaction list and transactions object

trans\_list <- split(country\_data$StockCode, country\_data$InvoiceNo)

transactions <- as(trans\_list, "transactions")

# Run Apriori

rules <- apriori(transactions, parameter = list(supp = 0.01, conf = 0.5, minlen = 2))

rules <- sort(rules, by = "lift", decreasing = TRUE)

# Print readable rules

if (length(rules) == 0) {

cat("No rules found for", country, "\n")

} else {

# Convert rules to data frame

rule\_df <- as(rules, "data.frame")

rule\_df$LHS\_Desc <- get\_descriptions(lhs(rules), stock\_map)

rule\_df$RHS\_Desc <- get\_descriptions(rhs(rules), stock\_map)

cat("\nTop 10 rules by lift for", country, ":\n")

print(rule\_df %>%

select(LHS\_Desc, RHS\_Desc, support, confidence, lift) %>%

arrange(desc(lift)) %>%

head(20))

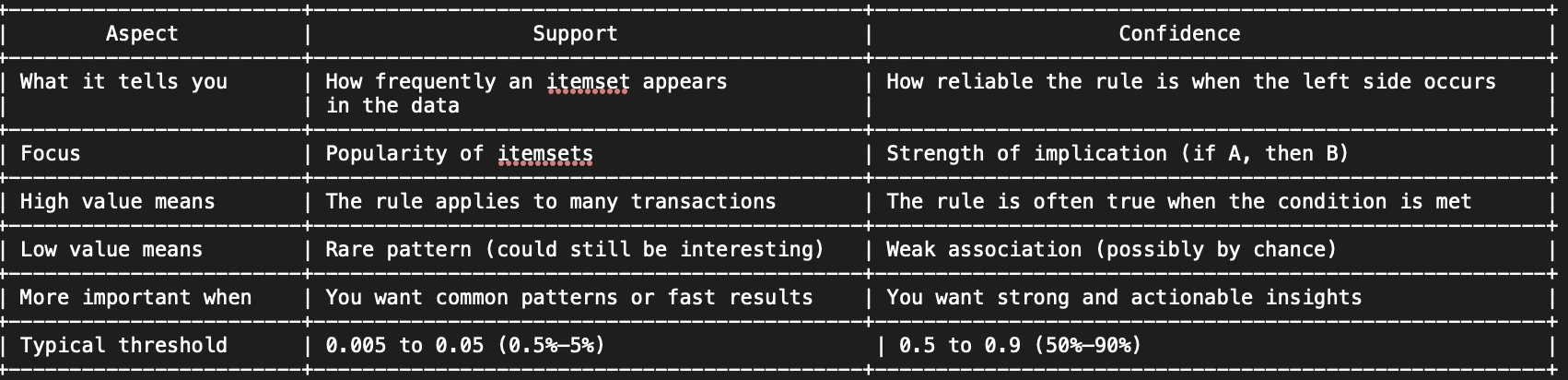
}

}

## Task 3 - Basket Analysis

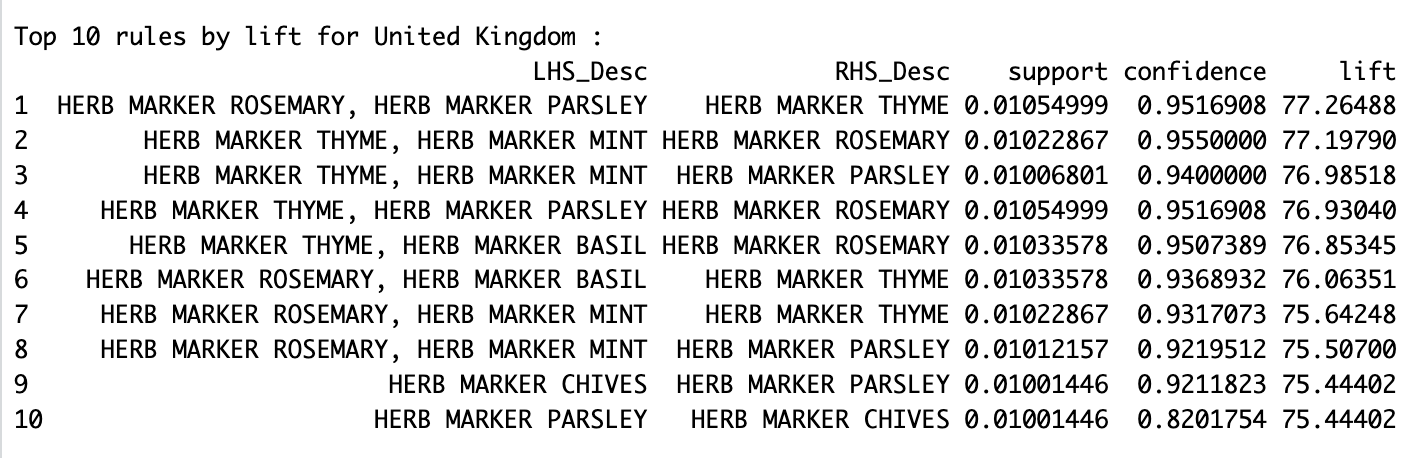
The initial Apriori run was done with a minimum support of 1%, confidence of 50%, and at least 2 items per rule. Although 70% - 80% will result strong reliable rules but it could lead to less useful rules.

Support vs Confidence



Market Basket Analysis for United Kingdom

Apriori top 10 rules for parameters supp = 0.01, conf = 0.5, minlen = 2



Rule interpretations

1. It seems Customers who buy both ROSEMARY and PARSLEY almost always buy THYME too. Ideal for bundling.

2. If someone buys THYME and MINT, they are very likely also buy ROSEMARY. These three may form a strong trio of complementary items.

3. Similar to Rule 2 MINT and THYME pair strongly with PARSLEY. A different trio than Rule 2, but equally reliable.

4. Shows ROSEMARY is often bought when people buy PARSLEY and THYME. Reinforces prior rule.

5. Even substituting MINT/PARSLEY with BASIL still results in strong ROSEMARY prediction. ROSEMARY is a key crossover item.

6. ROSEMARY and BASIL jointly predict THYME. These three consistently show up across rules.

7. ROSEMARY + MINT again leads to THYME — confirms THYME’s strong linkage with most other herb markers.

8. MINT + ROSEMARY also predict PARSLEY with very high confidence.

9. CHIVES buyers very likely get PARSLEY — high lift suggests strong bond, despite single-item

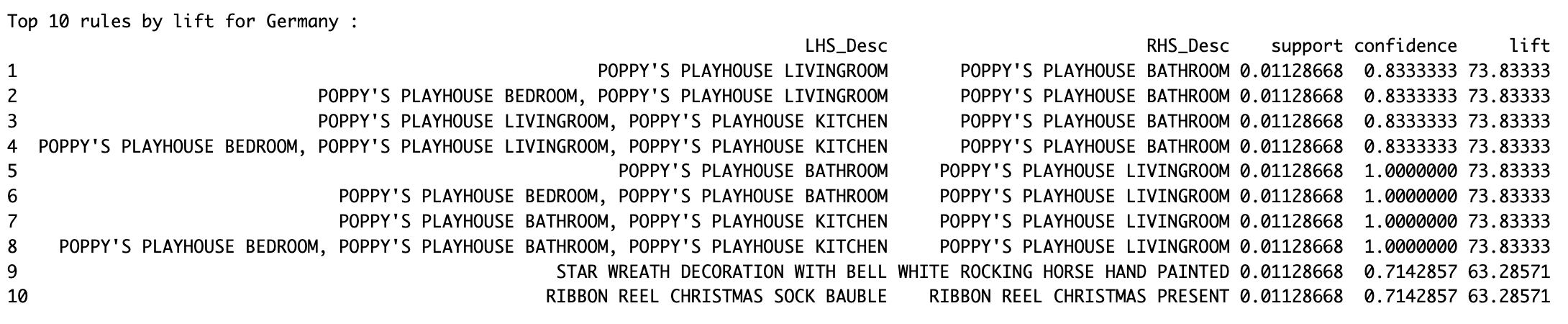
10. The reverse of Rule 9 — if someone buys PARSLEY, they frequently buy CHIVES too. Slightly lower confidence, same strong lift.

Overall, all these rules indicate extremely strong product associations within the “HERB MARKER” series of products. Almost all the rules’ projects confidence above 90% and lift above 75, which is exceptionally strong and incredibly high association strength. Seems like Herb markers are frequently purchased in combinations.

ROSEMARY, THYME, PARSLEY, and MINT consistently appear together, forming highly reliable combinations. Rules show that buying any two herbs almost guarantees purchase of a third. These findings are ideal for bundling and cross-selling.

Market Basket Analysis for Germany

Apriori top 10 rules for parameters supp = 0.01, conf = 0.5, minlen = 2 ordered for lift.



Overall, Germany shows strong associations within the "POPPY'S PLAYHOUSE" product line, especially between BATHROOM and LIVINGROOM sets.

Rules with 100% confidence suggest guaranteed purchases.

Holiday decorations like wreaths, baubles, and ribbon reels also exhibit strong bundling potential.

All 10 rules exhibit high lift values (63–74), indicating much higher co-occurrence.

Rule Interpretation

Rule 1 – 4

Customers who buy any playhouse room set almost always add the Bathroom set. Very strong cross-selling potential. With a confidence level of 83.33% and lift above 70% shows relatively strong and incredibly high association strength.

Rule 5–8:

With a confidence level of 100%, it seems Every customer who buys the BATHROOM set also buys the LIVINGROOM set. It's a guaranteed co-purchase highly valuable for promotions or recommendations.

Rule 9:

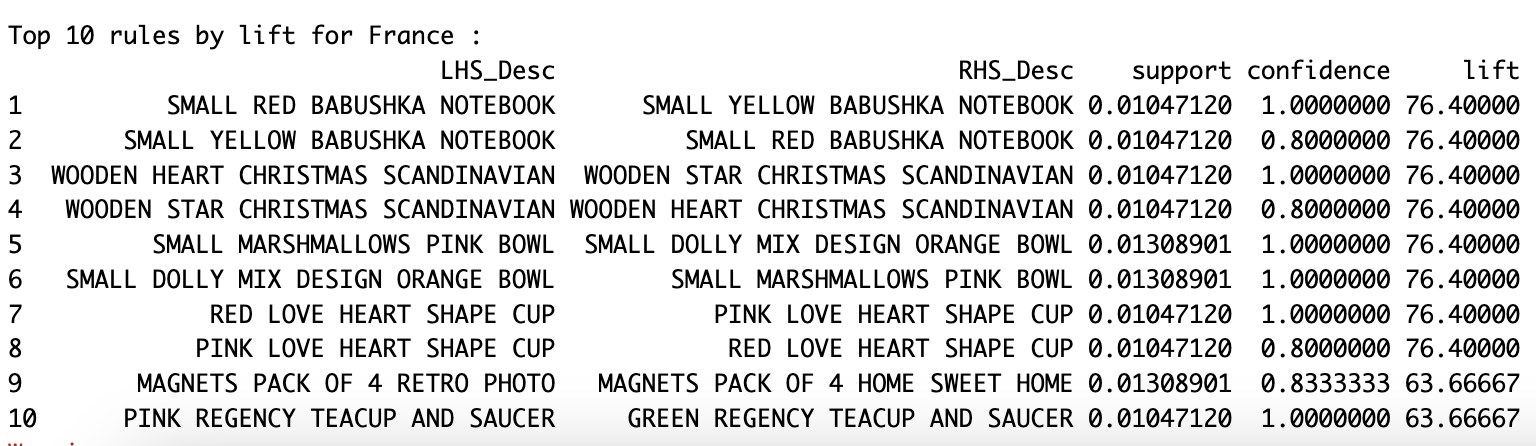
STAR WREATH DECORATION WITH BELL and WHITE ROCKING HORSE HAND PAINTED seems to be seasonal complementary items with strong association which are being bought together.

Rule 10

RIBBON REEL CHRISTMAS SOCK BAUBLE and RIBBON REEL CHRISTMAS PRESENT Christmas-themed ribbon items are highly associated. Strong opportunity for bundle deals during holiday seasons with a confidence level over 70%.

Market Basket Analysis for France

Apriori top 10 rules for parameters supp = 0.01, conf = 0.5, minlen = 2 ordered for lift.



Above rules depicts co-purchasing behavior among color or theme home variants. BABUSHKA notebooks, heart-shaped cups, and tea sets show 100% or near-perfect confidence, suggesting strong customer preference for matched or complementary products. These rules are ideal for bundling or gift-set curation.

Seasonal items like wooden Scandinavian decorations also pair reliably. Magnets and bowls reveal strong lift, reinforcing cross-sell value.

Rule Interpretation

Rule 1 – 2

SMALL RED BABUSHKA NOTEBOOK and SMALL YELLOW BABUSHKA NOTEBOOK seems to be highly correlated with a confidence value of 100% and lift of over 75%, therefore customers buying one BABUSHKA notebook very likely buy the other.

Rules 3 – 4

WOODEN HEART CHRISTMAS and WOODEN STAR CHRISTMAS highly correlated purchases with high confidence and lift values indicating frequent buys probably during the holidays. These two items can be seasonal and ideal to bundle together.

Rules 5 – 6

SMALL MARSHMALLOWS PINK BOWL and SMALL DOLLY MIX ORANGE BOWL seems to always be bought together which is ideal for bundling.

Rules 7 – 8

RED LOVE HEART CUP and PINK LOVE HEART CUP also being bought together just like above items. Mostly like as a pair or a set.

Rule 9

MAGNETS PACK OF 4 RETRO PHOTO likely to by HOME SWEET HOME product item.

Rule 10

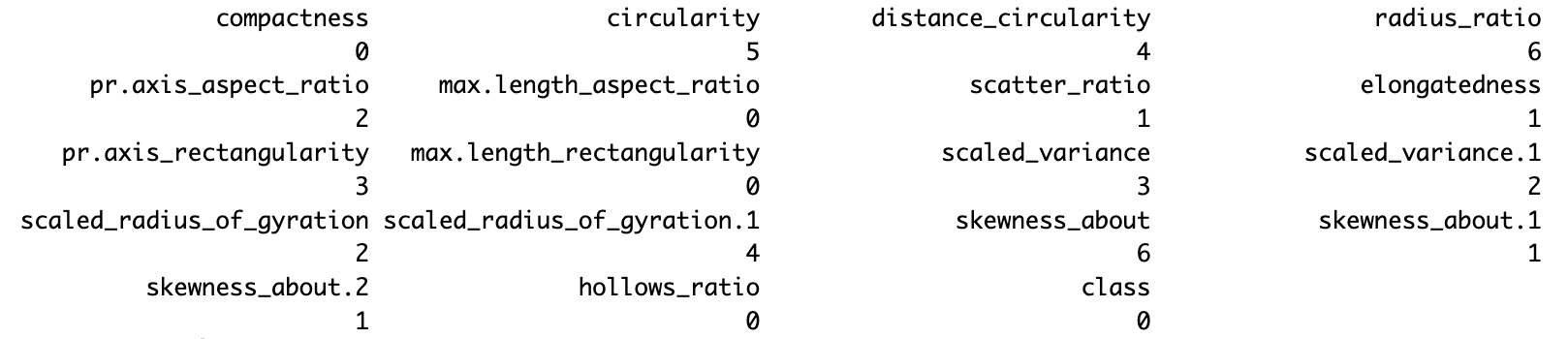
PINK REGENCY TEACUP AND SAUCER and GREEN REGENCY TEACUP AND SAUCER a guaranteed co-purchase with confidence level of 100%. Seems like customers buying the pink variant always buy the green one, suggesting strong interest in color-matched tea sets.

Question 2: Clustering

## Task 1. K-Means clustering after performing dimensionality reduction (PCA)

Following things were done to achieve to K Means clustering.

1. First data set is loaded to RStudio.
2. Verified that dataset contains all the columns, data types and then analyzed for missing values.



Some column had few missing values, and it was decided to **impute** these missing values rather than removing them from the dataset.

1. Then dataset was analyzed to find outliers, min, max and mean values.

A screenshot of a computer

AI-generated content may be incorrect.

Outlier counts

A screen shot of a computer code

AI-generated content may be incorrect.

Before Outlier Capping

A graph with red and black lines

AI-generated content may be incorrect.

It was decided to replace these outliers with a capping (winsorization) function. This will be beneficial as PCA is based on variance and covariance, and having outliers may orient principal components toward those outliers which could lead to misleading directions and poor clustering. Therefore, having outliers skewed toward a few distant points.

After Outlier Capped

A graph with blue and black lines

AI-generated content may be incorrect.

1. Data Analysis / Feature Correlation Analysis

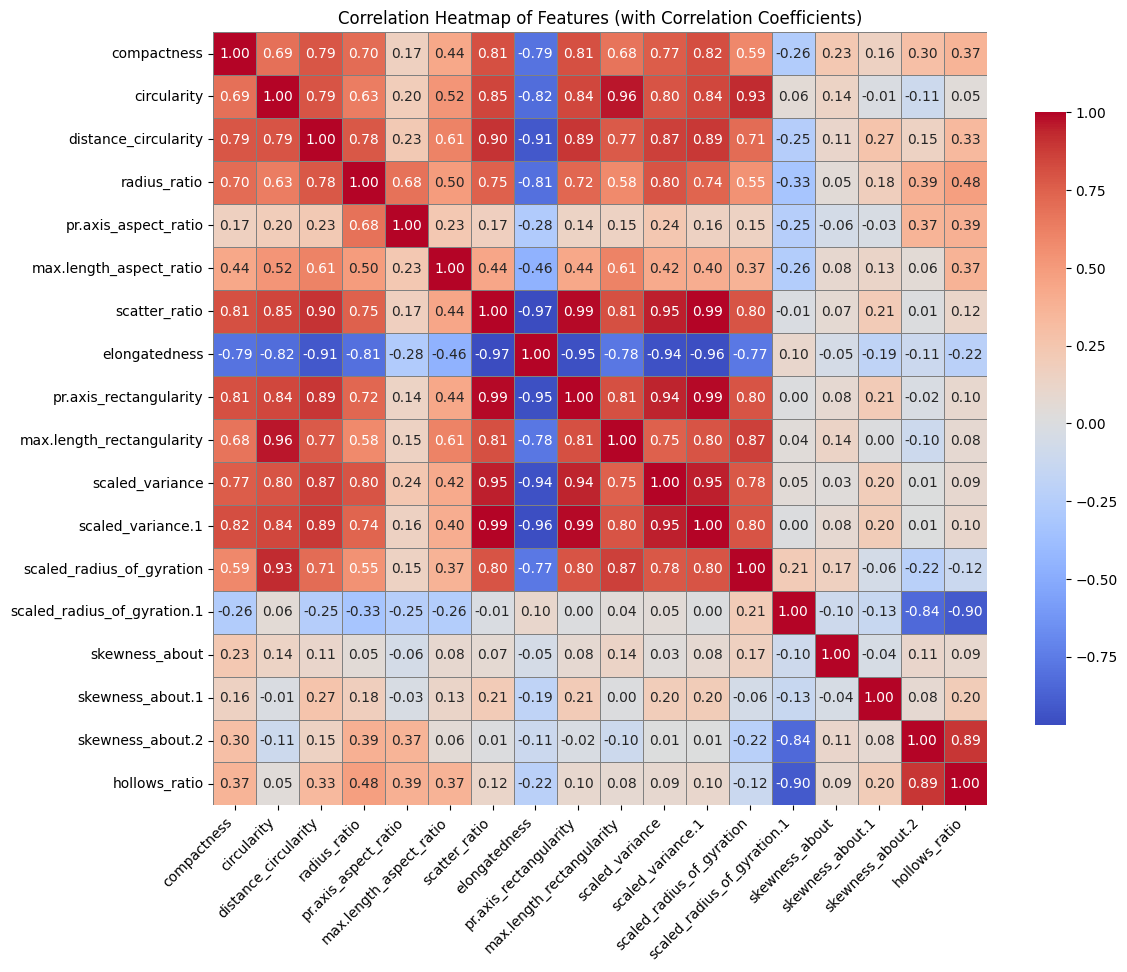
Univariate Analysis

A pie chart with numbers and a number of vehicles

AI-generated content may be incorrect.

 Pie Chart we indicates that approximately 51% are cars, 25% are buses.

Multivariate Analysis



The scatter\_ratio, elongatedness, and pr.axis\_rectangularity shows extremely high positive correlations with each other (exceeding +0.99). The scaled\_variance and scaled\_variance.1 are perfectly correlated suggesting feature duplication. From negative correlations, elongatedness shows strong inverse relationships with scaled\_variance.1 (–0.96), scaled\_variance (–0.95), and max.length\_rectangularity (–0.94).

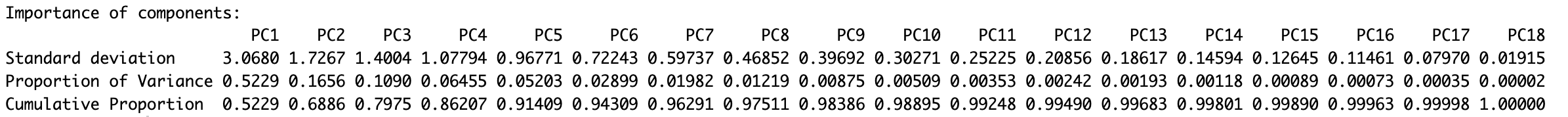
Pair Plot

A screenshot of a computer screen

AI-generated content may be incorrect.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | Car | Van | Bus | Observation |
| Compactness | High (~97) | Medium (~90) | Low (~89) | Cars show the highest compactness; buses the lowest |
| Distance Circularity | High (~94) | Medium (~75) | Low (~73) | Cars have high distance\_circularity |
| Radius Ratio | High (~186) | Low (~144) | Medium (~168) | Vans can be distinguished by low radius\_ratio |
| Circularity | Slightly High (~46) | Low (~42) | Medium (~44) | Helps distinguish cars when combined with other features |
| Pr.axis Aspect Ratio | Slightly High (~61) | Low (~59) | Medium (~64) |  |

1. After capping the outliers feature scaling done for standardization of columns. So PCA won't be biased to larger values and K-Means won’t treat large-scale features as more prominent.
2. Then dimensionality reduction was done using the prcomp() method.



A graph with numbers and lines

AI-generated content may be incorrect.

## According to the PCA results, PC1 alone captures over 52% of the variance, while PC1 through PC5 collectively account for more than 91% of the total variance in the dataset. This indicates that most of the structural information in the vehicle silhouette features can be effectively represented in a reduced 5-dimensional space. Therefore, dimensionality reduction to five principal components prior to clustering is both justified and effective, as it retains most of the original variance while simplifying the feature space.

1. Also, this further proves by observing the cumulative variance plot. Below cumulative variance plot, the black curve represents the cumulative variance, and the red dashed line represents the 90% threshold. The black curve crosses the red line at the 5th component indicating that the first five principal components capture over 90% of the total variance in the dataset.

A graph with a line and a red line

AI-generated content may be incorrect.

1. Then number of optimal clusters required for the K-means was found out that using the elbow method.

A graph with a line

AI-generated content may be incorrect.

According to the plot the most significant drop occurs between k = 1 and k = 4.

After k = 4, drop slows down. Between K= 4 to k = 6 drop is still there, but at a slower pace.

Therefore, the most noticeable "elbow" occurs around k = 4, K = 6 also acceptable if granular details are required.

So optimal clusters can be 4 or 6.

1. When 4 clusters were used.

PC1 and PC2 were chosen for plotting because they capture the most variance in the data (~69%).

A graph showing different colored dots

AI-generated content may be incorrect.

Cluster Profile

A white background with black text

AI-generated content may be incorrect.

Confusion Matrix

A number on a white background

AI-generated content may be incorrect.

### Conclusion

The vehicle dataset included properties such as scaled variance, skewness, kurtosis, compactness, circularity, rectangularity etc, and the vehicles include double-decker bus, Chevrolet van, and two car models: Saab and Opel Manta.

The experimental design anticipated that the bus, van, and one of the car types would be distinguishable, whereas the two car models might be more difficult to separate.

**Cluster 1 ( red):**

Cluster 1 has balanced feature values — compactness (92.2), radius ratio (175.0), and moderate circularity (44.8). These metrics suggest vehicles that don’t strongly exhibit car-, van-, or bus-like features. The red cluster lies near the center of the PCA plot and contains a mix of cars and vans. This aligns with the assignment's hypothesis that not all vehicle types would be cleanly separable, particularly the two car models (Saab and Opel Manta), which may appear in this blended group

**Cluster 2 (green):**

Cluster 2 shows moderate compactness (91.0) and lower circularity (38.9) with elevated radius ratio (161.0) and max.length rectangularity (135.0).

These features suggest more elongated, rectangular silhouettes, consistent with the geometry of Chevrolet vans. In the PCA plot, this group appears in the top-right green cluster, which is largely composed of square markers representing vans.

The clustering separates this group from both buses and cars, confirming the assignment’s assumption that vans would be distinguishable based on silhouette features.

**Cluster 3 (blue):**

Cluster 3 is the most compact and circular group, showing the highest compactness (104.0) and circularity (52.5) among all clusters. It also has the highest scatter ratio and the lowest elongatedness, indicating tightly packed, symmetric silhouettes.

Above traits align well with passenger car shapes, likely corresponding to either the Saab or Opel Manta. This cluster is visually identifiable in the far-left (blue) region of the PCA plot, where triangle markers (cars) dominate.

The high class purity and distinctive shape metrics make this cluster clearly distinguishable, supporting the experimental expectation that at least one car type would form a separate group.

**Cluster 4 (purple):**

Cluster 4 contains the lowest compactness (85.3), lowest radius ratio (129.0), and elevated elongatedness (48.6). These features point to larger, boxier vehicles — characteristic of buses. The purple cluster appears on the bottom-right of the PCA plot and includes many circle markers (buses), but also some overlap with cars and vans. This reflects the challenge of separating vehicles viewed from different angles, which may share similar aspect ratios or rectangularity. While still largely distinguishable, this cluster shows how viewpoint variation introduces some ambiguity, as expected in the experimental design.

## Task 2 - Agglomerative hierarchical clustering after performing dimensionality reduction (PCA)

Code : <https://github.com/lolitha-lakshan-1/DataMining_assignment_1/blob/main/Clustering.Rmd>

Code to perform Agglomerative hierarchical clustering

# Compute distance

# Euclidean distance reflects how far two points are from each other in terms of overall variance-base and data is continous

distance\_matrix <- dist(pca\_selected, method = "euclidean")

#Agglomerative Clustering

# Ward method works only with squared Euclidean distances

hc <- hclust(distance\_matrix, method = "ward.D2")

#Plot the Dendrogram

plot(hc, labels = FALSE, hang = -1, main = "Hierarchical Clustering Dendrogram")

#Cut Dendrogram into k Clusters

cluster\_cut <- cutree(hc, k = 4)

# Plot PCA

pca\_df <- as.data.frame(pca\_selected)

pca\_df$cluster <- factor(cluster\_cut)

pca\_df$class <- vehicles\_no\_outliers$class # Optional: for coloring by true class

library(ggplot2)

ggplot(pca\_df, aes(x = PC1, y = PC2, color = cluster, shape = class)) +

geom\_point(alpha = 0.7) +

labs(title = "Hierarchical Clustering (Ward) on PCA-Reduced Data",

x = "PC1", y = "PC2") +

theme\_minimal()

# Confusion Matrix

table(Cluster = pca\_df$cluster, Class = pca\_df$class)

Above code applies after reducing the dataset using PCA to five principal components and then euclidean distance matrix is computed to measure dissimilarity between observations.

Thereafter, Ward.D2 method is used to perform agglomerative clustering. A dendrogram is plotted to visualize how clusters merge at different distance levels. The tree is then cut at a chosen height to form four clusters.

A diagram of a clustering diagram

AI-generated content may be incorrect.

Above dendrogram depicts hierarchical structure in the PCA-reduced vehicle data. The biggest jump in height occurs around height 40–50, indicating 4 major clusters. This also proves the decision to use 4 clusters in K-Means as well.

The tree also shows that some vehicle shapes are grouped closely together in small subgroups, especially near the bottom of the plot.

A graph showing a number of colored dots

AI-generated content may be incorrect.

Confusion Matrix

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Cluster** | **Bus** | **Car** | **Van** | **Dominant Class** | **Explanation** |
| 1 | 71 | 131 | 107 | Mixed(Car Heavy) | Overlap between cars, vans, and buses due to similar shapes or views. |
| 2 | 23 | 227 | 1 | Car | Cluster consists almost entirely of cars. The shape features imply high compactness and symmetry, consistent with silhouettes of a single car model. |
| 3 | 88 | 71 | 91 | Mixed | This cluster holds a balanced mix of buses, vans, and cars. It likely captures complex silhouettes. This mixed group reflects the challenge of separating vehicles with similar profiles. |
| 4 | 36 | 0 | 0 | Bus | This cluster contains only buses. This clean separation supports the expectation that buses are distinguishable by shape. |

### Conclusion

## Cluster 1

Cluster 1 is positioned in the center-right area of the PCA plot and contains a heterogeneous mix of cars (131), vans (107), and buses (71). Vehicles in this group has moderate compactness (92.2) and mid-range circularity (44.8), suggesting moderately compact and symmetric silhouettes compared to other clusters. Their radius ratio (175.0) also lies between the elongated shapes of Cluster 2 and the more compressed buses in Cluster 4. The overlap in shape metrics across classes likely explains the class mixing seen here, highlighting the difficulty of clean separation for some vehicle views.

## Cluster 2

## Cluster 2 lies on the far-left side of the PCA plot and is dominated by cars (227), with minimal presence of buses (23) and vans (1). It features compactness (91.0) slightly lower than Cluster 1 but the lowest circularity (38.9) among the groups, suggesting less circular and more elongated shapes. While this might seem counterintuitive for cars, it could reflect side-profile silhouettes of one specific model that appears consistently in this region. The distinct PCA position indicate that this car group was well-captured by hierarchical clustering despite subtle geometric variations.

## Cluster 3

## Cluster 3 is found in the lower-right quadrant of the PCA plot, showing a balanced but mixed composition of buses (88), vans (91), and cars (71). With higher compactness (104.0) and circularity (52.5) than all other clusters, these vehicles appear most symmetric and rounded, likely representing silhouettes with frontal or rear views. Despite those strong shape signals, class labels are mixed, hinting that cars, vans, and buses may share similar outline structures from certain angles. This cluster captures the real-world ambiguity in silhouette data.

## Cluster 4

Cluster 4 appears in the bottom-left of the PCA plot and consists entirely of buses (36), making it the purest cluster. Vehicles in this group show the lowest compactness (85.3) and lowest radius ratio (129.0), indicating large, rectangular, and flattened shapes, likely representing double-decker bus silhouettes. Their elongatedness (48.6) and low circularity (40.7) reinforce this interpretation. The clean separation both visually and numerically supports the experimental hypothesis that buses are easily distinguishable based on shape features alone.

## Overall, the clustering results support the hypothesis that buses and vans are distinguishable based on shape, while separating car models remains more challenging.

## Task 3

The cluster analysis helps to identify vehicle categories such as buses, vans, and cars based on shape. This is important for camera-based classification systems, autonomous driving, smart traffic monitoring, and automated tolling where quick and reliable silhouette recognition is crucial.

The analysis also highlights outlier or rare shapes, enabling systems to flag them for special handling.

It also validates the assumption that buses and vans are more distinguishable than similar-looking cars, which can guide model training. Overall, this enhances the efficiency, scalability, and real-world usefulness of shape-based vehicle detection solutions.